Contents lists available at ScienceDirect

Heliyon

journal homepage: www.cell.com/heliyon

Research article

Quantitative structure-activity relationship model development for estimating the predicted No-effect concentration of petroleum hydrocarbon and derivatives in the ecological risk assessment

Jiajia Wei [a,b,c], Lei Tian [a,b,d], Fan Nie [a], Zhiguo Shao [a], Zhansheng Wang [a,**], Yu Xu [a], Mei He [a,b,c,*]

[a] State Key Laboratory of Petroleum Pollution Control, CNPC Research Institute of Safety and Environmental Technology Co., Ltd, Beijing, 102206, China
[b] Hubei Key Laboratory of Petroleum Geochemistry and Environment (Yangtze University), Wuhan, 430100, China
[c] School of Resources and Environment, Yangtze University, Wuhan, 430100, China
[d] School of Petroleum Engineering, Yangtze University, Wuhan, 430100, China

ARTICLE INFO

Keywords:
Petroleum hydrocarbons
Dietary exposure
Predicted No-effect concentrations
Quantitative structure-activity relationship
Toxicity

ABSTRACT

Quantitative structure-activity relationship (QSAR) is a cost-effective solution to directly and accurately estimating the environmental safety thresholds (ESTs) of pollutants in the ecological risk assessment due to the lack of toxicity data. In this study, QSAR models were developed for estimating the Predicted No-Effect Concentrations (PNECs) of petroleum hydrocarbons and their derivatives (PHDs) under dietary exposure, based on the quantified molecular descriptors and the obtained PNECs of 51 PHDs with given acute or chronic toxicity concentrations. Three high-reliable QSAR models were respectively developed for PHDs, aromatic hydrocarbons and their derivatives (AHDs), and alkanes, alkenes and their derivatives (ALKDs), with excellent fitting performance evidenced by high correlation coefficient (0.89–0.95) and low root mean square error (0.13–0.2 mg/kg), and high stability and predictive performance reflected by high internal and external verification coefficient (Q^2_{LOO}, 0.66–0.89; Q^2_{F1}, 0.62–0.78; Q^2_{F2}, 0.60–0.73). The investigated quantitative relationships between molecular structure and PNECs indicated that 18 autocorrelation descriptors, 3 information index descriptors, 4 barysz matrix descriptors, 6 burden modified eigenvalues descriptors, and 1 BCUT descriptor were important molecular descriptors affecting the PNECs of PHDs. The obtained results supported that PNECs of PHDs can be accurately estimated by the influencing molecular descriptors and the quantitative relationship from the developed QSAR models, that provided a new feasible solution for ESTs derivation in the ecological risk assessment.

1. Introduction

During the long-term exploration of oilfields and rapid development of petroleum industry, a wide variety of PHDs with high risks

[*] Corresponding author. State Key Laboratory of Petroleum Pollution Control, CNPC Research Institute of Safety and Environmental Technology Co., Ltd, Beijing, 102206, China.
[**] Corresponding author.
E-mail addresses: wangzs@cnpc.com.cn (Z. Wang), hemei-521@163.com (M. He).

https://doi.org/10.1016/j.heliyon.2024.e26808
Received 24 January 2024; Received in revised form 20 February 2024; Accepted 20 February 2024
Available online 28 February 2024

are produced and released into the adjacent environments [1–3]. It is estimated that the concentration of total petroleum hydrocarbons (TPHs) was 1.83 × 10^5 mg/kg in the soils and sediments around the productive oilfields [4,5], which exceeded the Soil environmental quality-Risk control standard of TPHs for development land (826 mg/kg) developed by Ministry of Ecology and Environment of China by 221 times [6]. A large number of PHDs are easily accumulated in organisms from the polluted environments, and pose potential high risks to high trophic organisms [7,8] and even human [9–11] through biomagnification [12], resulting in the decline of ecosystem services function [13] and the disrupt of the ecological mechanism [14]. However, accurate risk assessment for PHDs is limited due to the lack of environmental quality limits for specific PHDs. Present risk assessment of PHDs are mostly conducted based on the environmental quality and risk control standard of TPHs rather than specific PHDs.

Biotoxicity testing is widely-used for evaluating the environmental safety thresholds of pollutants [15], but usually limited by the labor-intensive, time-consuming, high-cost of the testing procedures and the raised ethical issues related to the animal testing [16,17]. QSAR models can act as a low-cost alternative for biotoxicity testing which directly estimate the biotoxicity based on the mathematical relationship between molecular structure and available toxicity concentrations [18,19]. In recent years, the advancements in computer technology also greatly promoted QSAR modelling strategies so that QSAR has been proposed as an effective technology for direct biotoxicity estimation by many authoritative environmental protection organizations such as REACH and OECD [20–22]. Previous studies have reported effective and reliable QSAR models for toxicity estimation of pesticides [18,23], 1,2,4-triazoles [24], pharmaceuticals and persistent organic pollutants [18], halogen derivatives, ethers and tertiary amines [25]. The molecular descriptors that describe the electrical, hydrophobicity, and thermodynamic structural characteristics, such as the energy of the highest unoccupied molecular orbital (E_{HOMO}), octanol-water partition coefficient ($logK_{ow}$), and overall or summation solute hydrogen bond basicity (MLFER_BO), were often found to be strongly correlated with the toxicity of chemicals [26–30]. However, little attention was paid to systematically investigate the quantitative relationship between the molecular structure and the biotoxicity and developed QSAR models for biotoxicity estimation of PHDs. Only a small amount of work has been devoted to develop QSAR models for estimating the acute toxicity of aromatic hydrocarbons and their derivatives (AHDs) in particular polycyclic aromatic hydrocarbon (PAHs) [19,29,31].

Current QSAR models are mostly developed for estimating the acute or chronic toxicity concentrations towards individual species, however, few QSAR models are directly developed to estimate the ESTs to the ecosystem. Little information are available to the ESTs of specific PHDs in the risk assessment, probably due to lack of sufficient toxicity data. Only ESTs for a few PAHs list as priority pollutants by U.S Environmental Protection Agency (U.S. EPA) [32–34] and aromatic hydrocarbons (AHs) [35] have so far been reported. PNECs is one of the ESTs that characterizes the magnitude of risks posed by the pollutants. The adverse effects of the pollutants are likely to

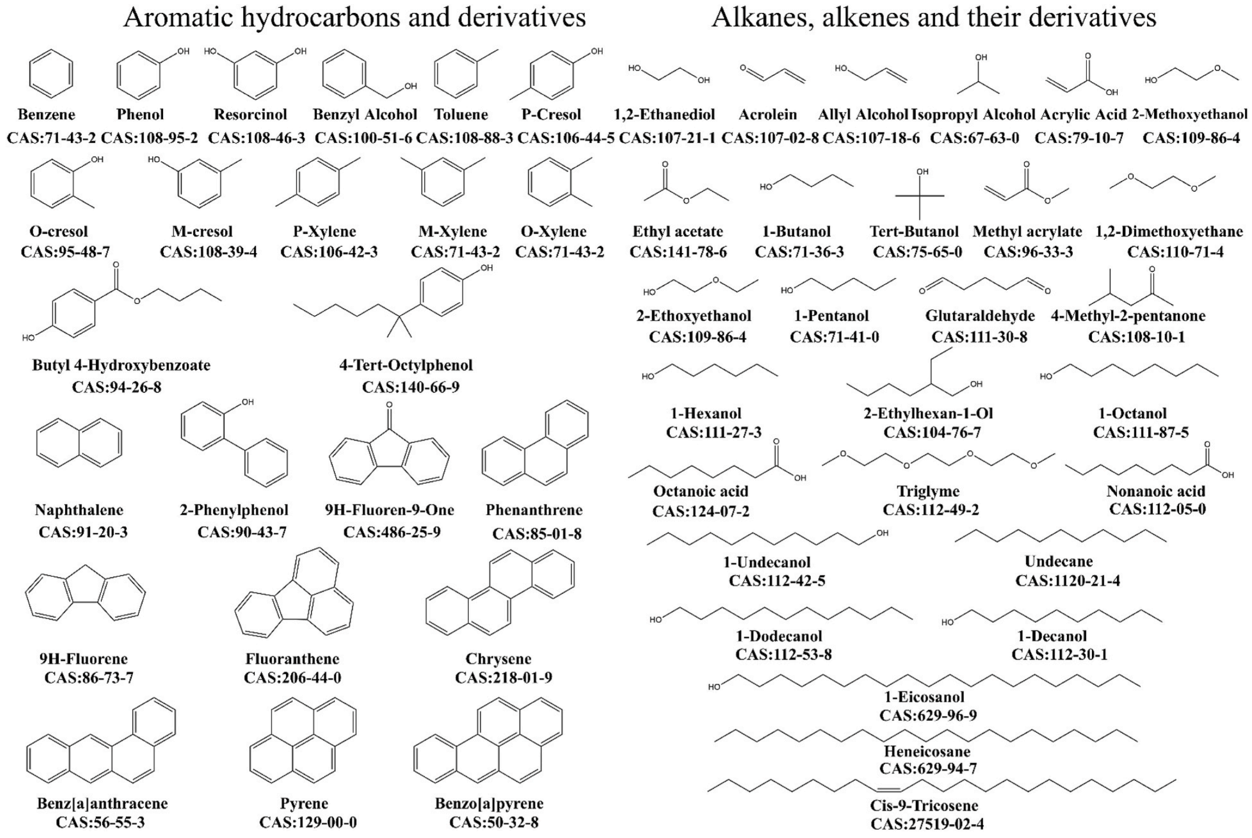

Aromatic hydrocarbons and derivatives
Alkanes, alkenes and their derivatives
Benzene CAS:71-43-2
Phenol CAS:108-95-2
Resorcinol CAS:108-46-3
Benzyl Alcohol CAS:100-51-6
Toluene CAS:108-88-3
P-Cresol CAS:106-44-5
1,2-Ethanediol CAS:107-21-1
Acrolein CAS:107-02-8
Allyl Alcohol CAS:107-18-6
Isopropyl Alcohol CAS:67-63-0
Acrylic Acid CAS:79-10-7
2-Methoxyethanol CAS:109-86-4
O-cresol CAS:95-48-7
M-cresol CAS:108-39-4
P-Xylene CAS:106-42-3
M-Xylene CAS:71-43-2
O-Xylene CAS:71-43-2
Ethyl acetate CAS:141-78-6
1-Butanol CAS:71-36-3
Tert-Butanol CAS:75-65-0
Methyl acrylate CAS:96-33-3
1,2-Dimethoxyethane CAS:110-71-4
2-Ethoxyethanol CAS:109-86-4
1-Pentanol CAS:71-41-0
Glutaraldehyde CAS:111-30-8
4-Methyl-2-pentanone CAS:108-10-1
Butyl 4-Hydroxybenzoate CAS:94-26-8
4-Tert-Octylphenol CAS:140-66-9
1-Hexanol CAS:111-27-3
2-Ethylhexan-1-Ol CAS:104-76-7
1-Octanol CAS:111-87-5
Naphthalene CAS:91-20-3
2-Phenylphenol CAS:90-43-7
9H-Fluoren-9-One CAS:486-25-9
Phenanthrene CAS:85-01-8
Octanoic acid CAS:124-07-2
Triglyme CAS:112-49-2
Nonanoic acid CAS:112-05-0
1-Undecanol CAS:112-42-5
Undecane CAS:1120-21-4
9H-Fluorene CAS:86-73-7
Fluoranthene CAS:206-44-0
Chrysene CAS:218-01-9
1-Dodecanol CAS:112-53-8
1-Decanol CAS:112-30-1
1-Eicosanol CAS:629-96-9
Heneicosane CAS:629-94-7
Benz[a]anthracene CAS:56-55-3
Pyrene CAS:129-00-0
Benzo[a]pyrene CAS:50-32-8
Cis-9-Tricosene CAS:27519-02-4

Fig. 1. The chemical structure of the petroleum hydrocarbons and derivatives.

occur when its exposure concentrations exceed the PNECs, especially after chronic or long-term exposure [36,37]. In this study, the quantitative relationship between PNECs and molecular structure was investigated and the QSAR models were developed for direct PNECs estimation of specific PHDs, which could greatly improve the ecological risk assessment of PHDs.

Dietary exposure is an important exposure route for PHDs from accidental ingestion, especially for the workers engaged in the oil industry, which can lead to severe bioconcentration and biomagnification for higher trophic organisms [38,39]. The present study focused on the risk assessment of PHDs from the dietary exposure and developed reliable QSAR models for the PNECs estimation of PHDs. The quantitative relationship between PNECs and molecular structure from the developed QSAR models was investigated to understand the underlying toxicity mechanisms of PHDs. The specific details are shown as follows: (1) All the existing acute and chronic toxicity concentrations for multiple toxicity endpoints of PHDs were collected from the US EPA-ECOTOX database and the current literatures; (2) All the collected toxicity concentrations were used for the PNECs derivation of PHDs, using the assessment factor (AF) approach [40–42]; (3) All the PHDs with the derived PNECs were selected as the specific PHDs datasets for QSAR model development; (4) The molecular structure of the selected specific PHDs were quantified by a series of molecular descriptors; (5) Reliable QSAR models were developed based on the PNECs and the molecular descriptors; (6) The quantitative relationship between the molecular structure and the PNECs from the developed QSAR models were investigated to understand the toxicity mechanisms of PHDs.

2. Materials and methods

2.1. Toxicity data collection and screening

All the existing toxicity concentrations (tested using OECD toxicity test methods) of PHDs (mainly AHDs and ALKDs) to plants, animals, and microorganisms through dietary exposure, including acute toxicity concentrations (e.g., median effective concentration EC_{50}, median lethal dose LD_{50}) and chronic toxicity concentrations (e.g., no observed effect concentration NOEC, lowest observed effect concentration LOEC) for multiple endpoints such as morphology, growth, histology, and reproduction, were collected from the US EPA-ECOTOX database (https://cfpub.epa.gov/ecotox/search.cfm) and the current literature. Detailed chemical information including chemical names, abbreviations, CAS numbers and chemical formulas of these PHDs were shown in Table S1 and their molecular structures were shown in Fig. 1.

The collected toxicity concentrations were firstly subjected to a preliminary screening. The toxicity concentrations that meet the following principles were selected: (1) the toxicity concentrations obtained using the toxicity testing methods proposed by the internationally recognized standard experimental guidelines; (2) the toxicity concentrations with clear exposure time and exposure route; (3) for chronic toxicity concentrations, NOEC and NOEL (no observed effect level) were preferred, but L (E)C_{10} (the concentration causing a 10% effect within a specified time interval) was also considered; (4) for the toxicity concentrations with a range, the minimum, mean, and maximum values were preferred. Then, the unit for all the selected toxicity concentrations was uniformly converted into mg/kg. The selected toxicity concentrations were used for subsequent PNECs estimation and QSAR model development.

2.2. PNECs estimation

PNECs were frequently employed as ESTs in the ecological risk assessment of pollutants for both aquatic and terrestrial ecosystems [43]. In this study, the PNECs of PHDs were estimated by dividing the acute toxicity concentrations (LD_{50}/EC_{50}) or chronic toxicity concentrations (NOEC/NOEL) by an appropriate assessment factor (AF), as described by Okonski et al. [44]. If all of the toxicity concentrations mentioned above is lacking, E (L) C_{10} or the half of LOEC or the half of LOEL (lowest observed effect level) was used instead of NOEC/NOEL, as recommended by the U.S. EPA [45]. AF was determined by the amount of the acute and chronic toxicity concentrations from different trophic levels (Table 1), according to the method described by Finizio et al. [37]. All the collected toxicity concentrations were used for the PNECs derivation of PHDs. PNECs of the PHDs with acute toxicity concentrations over three trophic levels or chronic toxicity concentrations over one trophic level were derived, according to the criteria requirements for PNECs derivation proposed by US EPA [45].

2.3. Quantification of molecular structure

The molecular structure of PHDs was quantified by multiple molecular descriptors in this study. The PHDs molecules were firstly visualized by ChemDraw 2D software and then optimized to their stable three-dimensional structures with the minimal energy by Chem3D software. The optimized molecular structure of PHDs was finally used to obtain a variety of molecular descriptors that

Table 1
The assessment factors used for deriving the PNECs of PHDs to the ecosystems.

Data set	Assessment factor
At least one short-term L(E)C_{50} from each of three taxonomic groups	1000
One long-term EC_{10} or NOEC from species representing one taxonomic group	100
Two long-term results (e.g., EC_{10} or NOECs) from species representing two taxonomic groups	50
Long-term results (e.g., EC_{10} or NOECs) from at least three species representing three taxonomic groups	10

describe different aspects of the molecular structure using PADEL software and ORCA software at the B3LYP/6-311G++ (d, p) level based on the Density Functional Theory. Octanol-Water Partition Coefficient ($LogK_{ow}$) of PHDs was characterized by EPIWEB4.1. As shown in Table S2 and 1444 two-dimensional molecular descriptors and 27 three-dimensional molecular descriptors, including 1 hydrophobic descriptor ($LogK_{ow}$), 17 electronic descriptors (e.g., E_{HOMO}, q_H^+, μ, α_{xx}), 1 steric descriptor (V_m), 8 thermodynamic descriptors (e.g., Eth, C_V, G^θ), 489 electrotopological state atom type descriptor (e.g., nHBint8, Shother, maxssssPb), 346 two-dimensional autocorrelation descriptors (e.g., AATSC0m, MATS1c, GATS1c), 96 burden modified eigenvalues descriptors (e.g., SpMax2_Bhm, SpMin6_Bhv), 91 barysz matrix descriptors (e.g., SpAbs_DzZ, SpMAD_Dzm, SpAbs_Dze), 67 ring count descriptors (e.g., nF8HeteroRing, n4HeteroRing, n3Ring), were obtained to characterize the molecular structure of PHDs.

2.4. *QSAR model development*

Before QSAR model development, a selection process was conducted to all the obtained molecular descriptors to avoid the overfitting in the QSAR modelling. The specific selection was conducted as follows: Firstly, the molecular descriptors with missing values were manually excluded. Then, the remaining molecular descriptors were imported into SPSS26 software to analyze the Pearson correlation coefficient between the molecular descriptors. Those molecular descriptors with high correlations that an absolute value of Pearson correlation coefficient was higher than 0.95 were removed to eliminate multicollinearity [46,47]. After this selection process, a total of 488 molecular descriptors were left for subsequent QSAR modelling.

The QSAR model was developed with -logPNEC as the dependent variable and molecular descriptors as the independent variables, using multiple linear regression (MLR) by SPSS26 software, according to the OECD QSAR guidelines “an unambiguous algorithm”. MLR was performed stepwise until passing the tests ($P < 0.05$) and identified the most important molecular descriptors for -logPNEC. Based on the results of the stepwise regression, a preliminary QSAR model was developed. High reliable QSAR models are characterized by higher adjusted multiple correlation coefficient ($R^2 > 0.6$) and the multicollinearity diagnosis (variance inflation factor, VIF<10). Ordinary least squares were used to eliminate insignificant molecular descriptors using *F*-test and *t*-test. If the *F*-test and *t*-test did not pass ($P > 0.05$) or if R^2 was small (<0.6), the regression were re-performed [48,49]. If VIF>10, the principal components of the variables were extracted and the regression analyses were re-performed to eliminate covariances. The specific formulas for the above parameters are shown in Table S3.

Double cross validation (Internal and external validations) was performed to the preliminary QSAR model, according to the OECD QSAR guidelines “appropriate measures of goodness-of-fit, robustness and predictivity” [49]. The training sets and validation sets were randomly selected in an approximate ratio of 4:1 from all the PHDs dataset, AHDs datasets and ALKDs datasets, respectively. Then, the training set is internally validated by leave-one-out (LOO) cross-validation to assess the internal robustness of the separate three QSAR models for PHDs, AHDs and ALKDs. The internal verification coefficient Q^2_{LOO} exceeding 0.5 indicates the developed QSAR model with good robustness [50,51]. Q^2_{LOO} is calculated using formula (1). The external validation coefficients Q^2_{F1} and Q^2_{F2} of the validation set were utilized to assess the predictive performance of the model. Both of Q^2_{F1} and Q^2_{F2} exceed 0.5 indicates the developed QSAR model with good external prediction ability [51]. Q^2_{F1} and Q^2_{F2} are calculated using formulas (2) and (3), respectively.

$$Q^2_{LOO} = 1\text{-}\frac{\sum_{i=1}^{n_{training}} (y_i^{exp} - y_i^{pred}\widehat{)}\,2}{\sum_{i=1}^{n_{training}} (y_i^{exp} - \bar{y}\widehat{)}\,2} \tag{1}$$

where y^{exp} and y^{pred} are the estimated and predicted -logPNEC of the training set, $\bar{y}$ is the average -log PNEC concentration of the training set, $n_{training}$ is the chemical number of the training set.

$$Q^2{}_{F1} = 1\text{-}\frac{\sum_{i=1}^{n_{test}} (y_i^{exp} - y_i^{pred}\widehat{)}\,2}{\sum_{i=1}^{n_{test}} (y_i^{exp} - \overline{y_{training}}\widehat{)}\,2} \tag{2}$$

$$Q^2{}_{F2} = 1\text{-}\frac{\sum_{i=1}^{n_{test}} (y_i^{exp} - y_i^{pred}\widehat{)}\,2}{\sum_{i=1}^{n_{test}} (y_i^{exp} - \overline{y_{test}}\widehat{)}\,2} \tag{3}$$

where y^{exp} and y^{pred} are the estimated and predicted -logPNEC of the validation set, $\overline{y_{training}}$ and$\overline{y_{test}}$ are the average -log PNEC of the training set and the validation set, n_{test} is the chemical number of the validation set.

The QSAR models passed the internal and external validations were developed for the PNECs estimation in this study. The -logPNEC (L) is described with the optimal combination of influential molecular descriptors (X_1, X_2 … X_n) used as independent variables. formula (4) is represented as follows:

$$L = K_1 \cdot X_1 + K_2 \cdot X_2 + \ldots + K_n \cdot X_n + K_0 \tag{4}$$

where L represents the dependent variable -logPNEC, X_1, …, X_n denote the independent variables of the molecular descriptors, K_1, …, K_n are the unstandardized coefficients of the independent variables, and k_0 is the constant term.

2.5. *Application domain analysis of the QSAR model*

In general, QSAR model development has its own limitations due to some influencing factors such as the sample number restriction and the model algorithm [52]. The application domain (AD), defined as a threshold for the chemicals that can apply the developed models, is usually analyzed through outlier detection. In this study, a well-defined application domain range for the QSAR model was analyzed through the identification of structural outliers and predicted outliers. It is not reliable to use the developed QSAR model with chemicals outside the application domain to estimate the PNECs. The chemicals with structural outliers in the training set and validation set were identified by the hat value (h) using the leverage approach [53], following formula (5). The warning leverage (h*) was calculated as formula (6). The chemicals with a hat value (h) higher than warning leverage (h*) were considered as structural outliers.

$$h = x_i \left(X^T X\right)^{-1} X_i^T \tag{5}$$

$$h^* = 3(k+1)/n \tag{6}$$

where x_i is the molecular descriptor of the i_{th} chemical, X represents the matrices of molecular descriptors, k is the number of molecular descriptors, n is the number of chemicals in the training set.

The chemicals with predicted outliers were identified according to the standardized residuals between the estimated and predicted -log PNEC of the chemicals. The standardized residual (δ) was calculated as formula (7). The chemicals that exceed the threshold of the standardized residuals (from −3 to 3) were considered as predicted outliers [46,49].

$$\delta = \frac{y - y^{pred}}{\sqrt{\frac{\sum_{i=1}^{n}(y-y^{pred}\widehat{)\;2}}{n-k-1}}} \tag{7}$$

where y and y^{pred} are the estimated and predicted -log PNEC of the training set and validation set, n is the number of chemicals in the training set, k is the number of molecular descriptors.

Then, the standardized residuals versus hat values of the chemicals in both the training set and validation set were plotted to visualize the outliers and establish the application domain range for the developed the QSAR models, according to the distance-based methods [54,55].

2.6. Quantitative relationship between PNECs and molecular structure

The quantitative relationship between PNECs and molecular structure was demonstrated by the standardized coefficients of the molecular descriptors involved in the developed QSAR models. The standardized coefficients of the molecular descriptors indicated the influencing weight of its effect on PNECs and were used to describe the quantitative relationship between PNECs and molecular structure. The specific details were conducted as follows. Firstly, $K_1, \ldots, K_n$ in formula (4), the unstandardized coefficients for the 1_{st}-n_{th} influencing molecular descriptors, were transferred to the corresponding standardized coefficients of the molecular descriptors

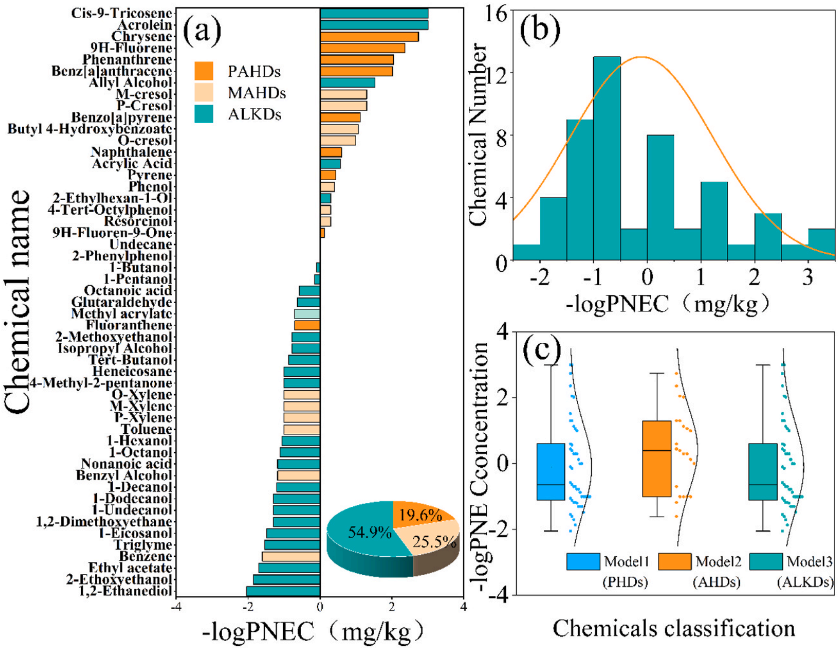


Fig. 2. (a) The -logPNEC concentrations of different PHDs; (b) The distribution of the -logPNEC concentrations of the PHDs for the three QSAR models; (c) The -logPNEC concentrations of the PHDs in the individual QSAR model.

Table 2
Biological species involved in the PNECs estimation and the estimated PNECs.

PHDs	Chemical name	Acute toxicity		Chronic toxicity		Assessment Factor	PNECs (mg/kg)
		Aquatic	Terrestrial	Aquatic	Terrestrial		
PAHDs	Naphthalene		*Colinus virginianus*		*Colinus virginianus, Mus musculus, Oryctolagus cuniculus, Rattus norvegicus*	10	0.25
	2-Phenylphenol		*Anas platyrhynchos, Colinus virginianus*		*Rattus norvegicus, Anas platyrhynchos, Colinus virginianus*	100	1
	9H-Fluoren-9-One		*Rattus norvegicus*		*Rattus norvegicus*	100	0.75
	Phenanthrene	*Neanthes arenaceodentata*		*Neanthes arenaceodentata, Platichthys flesus*	*Porcellio scaber, Mus musculus, Mesocricetus auratus, Rattus norvegicus*	10	0.0089
	9H-Fluorene				*Oniscus asellus, Rattus norvegicus*	50	0.0044
	Fluoranthene			*Nereis virens,*	*Porcellio scaber, Rattus norvegicus, Mus musculus*	10	5
	Chrysene			*Platichthys flesus*	*Drosophila melanogaster*	50	0.0018
	Benz [*a*] anthracene				*Oniscus asellus, Porcellio scaber*	100	0.0096
	Pyrene		*Acheta domesticus,*		*Orchesella cincta, Mus musculus, Rattus norvegicus*	50	0.35
	Benzo [*a*]pyrene			*Fundulus heteroclitus,*	*Oniscus asellus, Rattus norvegicus, Gallus gallus,*	10	0.074
MAHDs	Benzene				*Drosophila melanogaster, Mus musculus*	50	40
	Phenol	*Oncorhynchus mykiss*			*Rattus norvegicus*	100	0.4
	Resorcinol				*Mus musculus, Rattus norvegicus*	100	0.5
	Benzyl Alcohol				*Drosophila melanogaster, Mus musculus*	50	25
	Toluene				*Mus musculus*	100	10
	P-Cresol		*Mouse,Rat*		*Drosophila melanogaster, Mus musculus, Oryctolagus cuniculus, Rattus norvegicus*	100	0.05
	O-cresol		*Mustela putorius, Neovison vison*		*Drosophila melanogaster, Mustela putorius, Neovison vison*	50	0.1
	M-cresol				*Oryctolagus cuniculus, Rattus norvegicus*	100	0.05
	P-Xylene				*Rattus norvegicus*	100	10
	M-Xylene				*Rattus norvegicus*	100	10
	O-Xylene				*Rattus norvegicus*	100	10
	Butyl 4-Hydroxybenzoate	*Oncorhynchus mykiss*		*Oncorhynchus mykiss*	*Mus musculus*	50	0.086
	4-Tert-Octylphenol	*Oryzias latipes*		*Platichthys flesus*		100	0.5
ALKDs	1,2-Ethanediol				*Mus musculus*	100	110.9
	Acrolein		*Anas platyrhynchos, Colinus virginianus, Mus musculus*		*Mus musculus, Oryctolagus cuniculus, Rattus norvegicus, Canis familiaris*	50	0.001
	Allyl Alcohol				*Mus musculus, Rattus norvegicus*	100	0.03
	Isopropyl Alcohol				*Rattus norvegicus*	100	6.01
	Acrylic Acid				*Rattus norvegicus*	100	0.27
	2-Methoxyethanol				*Drosophila melanogaster, Mus musculus, Rattus norvegicus*	50	6
	Ethyl acetate				*Ostrinia nubilalis*	100	50

(*continued on next page*)

Table 2 (*continued*)

PHDs	Chemical name	Acute toxicity		Chronic toxicity		Assessment Factor	PNECs (mg/kg)
		Aquatic	Terrestrial	Aquatic	Terrestrial		
	1-Butanol		*Mouse, Hamster, Bird, Dog*		*Rattus norvegicus*	100	1.25
	Tert-Butanol				*Mus musculus, Rattus norvegicus*	100	7.41
	Methyl acrylate				*Drosophila melanogaster*	100	5
	1,2-Dimethoxyethane		*Mus musculus*		*Mus musculus*	100	20
	2-Ethoxyethanol				*Drosophila melanogaster, Mus musculus*	50	72.1
	1-Pentanol				*Rattus norvegicus*	100	1.4
	Glutaraldehyde		*Anas platyrhynchos, Colinus virginianus*		*Drosophila melanogaster, Anas platyrhynchos, Colinus virginianus*	50	4.3
	4-Methyl-2-pentanone		*Rattus norvegicus*		*Rattus norvegicus*	100	10
	1-Hexanol		Rat		Rat	100	11.27
	2-Ethylhexan-1-Ol		*Mice, Rabbits, Guinea pigs, Rattus norvegicus*		*Rattus norvegicus*	100	0.5
	1-Octanol		*Rattus norvegicus*		*Rattus norvegicus*	100	13
	Octanoic acid		*Heterobothrium okamotoi,*	*Pagrus major, Takifugu rubripes*	*Rattus norvegicus*	10	3.75
	Triglyme		*Mus musculus*		*Mus musculus*	100	35
	Nonanoic acid		*Mice, Colinus virginianus*		*Rattus norvegicus*	100	15
	1-Undecanol		*Rattus norvegicus*		Male rats	100	20
	Undecane				*Rattus norvegicus*	100	1
	1-Dodecanol				*Rattus norvegicus, Male rats*	100	20
	1-Decanol		*Rattus norvegicus*		*Rattus norvegicus*	100	15.83
	1-Eicosanol				*Rattus norvegicus*	100	29.85
	Heneicosane				*Wistar rats*	100	10
	Cis-9-Tricosene		*Colinus virginianus, Anas platyrhynchos, Oryctolagus cuniculus, Rattus norvegicus*		*Anas platyrhynchos*	100	0.001

according to formula (8).

$$K_i^* = K_i/(S_L/SX_i) \tag{8}$$

where K_i^* is the standardized coefficient of a molecular descriptor, K_i is the unstandardized coefficient of the molecular descriptor, S_L is the standard deviation of the dependent variable L, and SX_i is the standard deviation of the independent variable X_i.

Then, the influencing weight of the molecular descriptors on the PNECs (W_i) were calculated using formula (9), based on their standardized coefficients.

$$W_i\ (\%) = K_i^*/\ (K_1^* + K_2^* + \ldots + K_n^*)\ ^*100\% \tag{9}$$

where K_1^*, …,K_n^* represent the standardized coefficient of the 1_{st}-n_{th} molecular descriptor.

3. Results and discussion

3.1. The molecular structural information of PHDs

In this study, the molecular structure of PHDs used for QSAR model development was shown in Fig. 1 and the additional detailed chemical information was provided in Table S1. The PHDs were composed of 45.1% AHDs, 47.1% alkenes and derivatives, and 7.8% alkanes and derivatives, indicating the 51 PHDs selected in this study covered a wide range of molecular structures (Fig. 2a). Multiple molecular descriptors, including hydrophobic descriptors, electronic descriptors, steric descriptors, two-dimensional autocorrelation descriptors, and information index descriptors (Table S2), were used to characterize different aspects of the molecular structural information of these PHDs and provide a detailed description of the molecular structural features. The large variations in the molecular descriptors also supported that the 51 PHDs selected in this study covered a wide variety of diverse molecular structures with quite different molecular structural properties (Tables S4–S7). For example, AHDs showed a stronger ability to gain and lose electrons than ALKDs, as reflected by higher q^- (a electronic descriptor that described the ability of chemicals to gain or lose electrons [56]) of AHDs (ranged from −0.16 to −2.04) than ALKDs. AHDs has a stronger hydrophobicity than ALKDs, as evidenced by higher MLOGP (a two-dimensional molecular descriptor that described the chemical hydrophobicity [57]) of AHDs (ranged from 1.9 to 3.66) than ALKDs. The variation of the information index descriptors (e.g., IC0–IC4, TIC0–TIC4, SIC0–SIC4, CIC0–CIC5, BIC0–BIC4, MIC0–MIC4, ZMIC0–ZMIC4) of AHDs (changed from 1.53 to 12.4 times) was also much larger than that of ALKDs (varied from 1.70 to 31.4 times).

3.2. The PNECs of PHDs

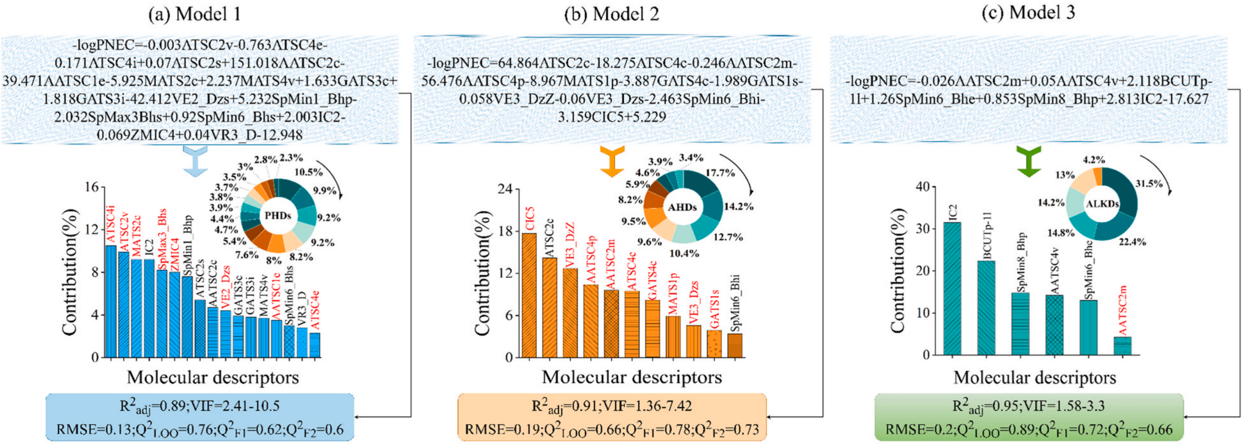
As presented in Table 2, the PNECs of 51 PHDs were obtained based on the acute or chronic toxicity concentrations (e.g., LD_{50}, NOEC) of PHDs to various vertebrates (e.g., rodents, rabbits, pigs, fish, and reptiles) and invertebrates (e.g., worms, arthropod). The toxicity concentrations specifically used for the PNECs estimation, were summarized in detail in Tables S8–S9. As the PNECs visualized in Table 2, the -logPNEC of the PHDs exhibit a normal distribution (Fig. 2b) and differed by nearly five logarithmic units (range from −2.04 to 3, Fig. 2a), indicating a large difference in the toxicity among these PHDs. Thus, the QSAR models based on these representative PHDs with wide-range toxicity in this study can be better applied to the toxicity estimation for diverse PHDs.

The results showed that the PNECs varied significantly (0.001–110.9 mg/L) with the type of the PHDs, following the order of polycyclic aromatic hydrocarbons and derivatives (PAHDs) > monocyclic aromatic hydrocarbons and derivatives (MAHDs) > ALKDs. The PNECs of AHDs were significantly lower than that of ALKDs, indicating higher toxicity of AHDs than ALKDs, which was in agreement with the toxicity investigation of aromatic hydrocarbons and long-chain n-alkanes [58]. In this study, there were 16 AHDs and 6 ALKDs in the 22 PHDs of higher toxicity characterized with -logPNEC>0, whereas 7 AHDs and 22 ALKDs in the 29 PHDs of lower toxicity characterized with -logPNEC<0, indicating larger proportion of the investigated AHDs with higher toxicity and lower proportion of the studied ALKDs with lower toxicity (Fig. 2a). Among the AHDs, the PNECs of PAHDs (ranged from 0.001824 to 5) were significantly lower than that of MAHDs (ranged from 0.05 to 40), indicating a higher toxicity of PAHDs (Table 2), which was consistent with previous finding of the hazardous concentration for 5% of species (HC_5) for AHDs [59].

3.3. The developed QSAR models

In the present study, three QSAR models were separately developed for all the PHDs (model 1, 51 datasets), the AHDs (model 2, 23 datasets), and the ALKDs (model 3, 28 datasets), strictly following the procedures of the OECD QSAR guidelines. The model equations and the performance of the three QSAR models were shown in Fig. 3, and the molecular descriptor descriptions were shown in Table S10. Results showed that all the three models showed good fitting performance, as evidenced by their high goodness of fit (high R^2 (0.89,0.91,0.95) and low RMSE values (0.13,0.19,0.2)). The closer R^2 is to 1, the better the goodness of fitting for the developed QSAR model.The three models are internally robust and stable, as reflected by the internal validation parameters (high Q^2_{LOO} (0.76,0.66,0.89)), and showed excellent external prediction capability by the high external validation coefficients (Q^2_{F1}: 0.62,0.78,0.72; Q^2_{F2}: 0.6,0.73,0.66). All of the model parameters were much higher than the acceptable thresholds of the OECD QSAR development requirements ($R^2 > 0.6$; $Q^2{}_{LOO} > 0.5$; $Q^2_{F1} > 0.5$, $Q^2_{F2} > 0.5$) [48,51].

The predictive performance of the developed QSAR models were visualized by the comparison of the experimental and predicted -logPNEC of the PHDs in the training set and validation set in Fig. 4. The results showed that the R^2 of the regression line fit for the three models were high to 0.93, 0.96 and 0.96, respectively, indicating a high degree of fitting for the developed models. Both of the experimental and the predicted -logPNEC in both the training set and validation set were evenly distributed on both sides of the regression line, and were in very good agreement with each other, supporting a high prediction accuracy of these models. Relatively small residuals of the predicted values against the experimental values of -logPNEC were observed for the three models, which were 0.01–1.11, 0.02–0.83, 0–0.96, respectively (Table S11), indicating the developed three models without systematic errors.



NOTE:Red labelled descriptors are positively correlated with logPNEC, black labelled descriptors are negatively correlated with logPNEC

Fig. 3. The model equations and parameters of the three developed QSAR models and the influencing weight of the molecular descriptors on the PNECs of PHDs.

Comparatively, the sum of squared residuals in the model 1 (6.38) was much larger than model 2 (2.563), suggesting model 2 is the more accurate for estimating the PNECs of AHDs. Similarly, the sum of squared residuals in the model 1 (3.82) was slightly larger than using model 3 (3.40), suggesting model 3 was better for the PNECs estimation of ALKDs with higher accuracy.

The application domain range of the developed QSAR models was defined and visualized via a Williams plot (h_i < h*, −3< standardized residuals<3) based on the standardized residuals versus hat values (h) of the PHDs in both the training set and the validation set (Fig. 4). The h values of the PHDs involved in both the training set and the validation set were below their respective waring leverage (h* = 1.256, 1.895 and 0.955) of the three models, indicating no structural outliers of PHDs existed in these models. The standardized residuals for all the PHDs involved in the three models did not exceed the standardized residual threshold (from −3 to 3), indicating no predicted outliers of PHDs in all the developed models. Therefore, all the PHDs involved in the developed three models are within the application domain. It is reliable to use the three QSAR models to estimate the PNECs of PHDs, AHDs, and ALKDs, respectively.

3.4. The QSAR model accuracy in the PNEC estimation

In this study, model 2 and model 3 were suggested to estimate the PNECs of AHDs and ALKDs, respectively. Three aromatic hydrocarbons (Naphthalene, 1-Methylnaphthalene, and Pyrene) and two alkanes (*n*-Decane and *n*-Heptane) that within the application domain range in the application domain analysis (Fig. 5a) and not included in the previous QSAR modeling were used to verify the

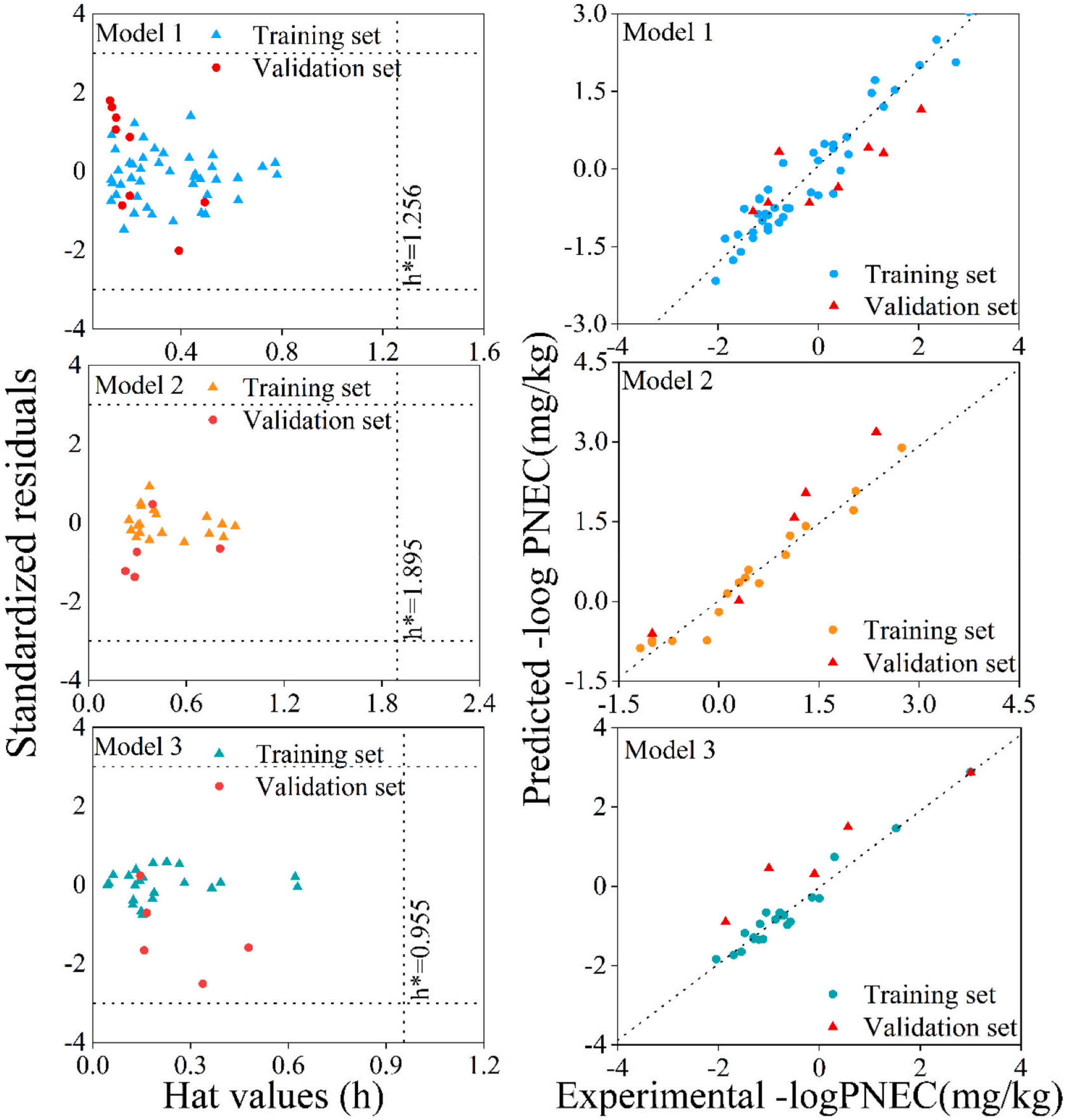


Fig. 4. The application domain range analysis and the predictive performance of the three developed QSAR models.

accuracy of the two models by comparing the estimated PNECs using the developed models with the published regulatory limits of these PHDs by international authoritative environmental protection organizations. The results showed that the estimated PNECs of Naphthalene, 1-Methylnaphthalene, and Pyrene (0.455, 0.044, and 0.255 mg/kg) were significantly lower than that for *n*-Decane and *n*-Heptane (2.771 and 4.70 mg/kg). The estimation results for the PNECs of the three aromatic hydrocarbons and two alkanes were consistent with their toxicity. As shown in Fig. 5b, the estimated PNEC of naphthalene was then compared with the peer-reviewed toxicity concentrations published by the European Food Safety Authority (EFSA), and the estimated PNECs of 1-Methylnaphthalene, Pyrene, *n*-Decane and *n*-Heptane were compared with the proposed safety limits published by the Environmental Protection Agency (EPA). The PNECs of these PHDs estimated by the developed models were approximate to the published regulatory limits by 0.07 – 0.36 log units. The obtained results supported that the developed models were with high accuracy in PNECs estimation for PHDs with diverse molecular structures.

3.5. *The quantitative structure-PNECs relationship*

The quantitative relationships between the molecular structure and PNECs in the present study were directly obtained from the developed QSAR models (Fig. 4). A total of 17, 11 and 6 molecular descriptors were related to the PNECs of all the PHDs (model 1), the AHDs (model 2), and the ALKDs (model 3), respectively. The influence of the molecular descriptors on the logPNEC varied with PHDs. For all the PHDs, eight molecular descriptors (ATSC2v, ATSC4e, ATSC4i, AATSC1e, MATS2c, VE2_D2S, SpMax3_Bhs, and ZMIC4) were positively correlated with the PNECs, whereas nine molecular descriptors (ATSC2s, AATSC2c, MATS4v, GATS3c, GATS3i, SpMin1_Bhp, SpMin6_Bhs, IC2 and VR3_D) were negatively correlated with the PNECs. However, for the AHDs, nine molecular descriptors (ATSC4c, AATSC2m, AATSC4p, MATS1p, GATS4c, GATS1s, VE3_DzZ, VE3_Dzs and CIC5) were observed to be positively with the logPNEC. Two descriptors (ATSC2c and SpMin6_Bhi) were negatively related with the logPNEC of the AHDs. For the ALKDs, the molecular descriptor AATSC2m was positively correlated with the PNECs, and five molecular descriptors (AATSC4v, BCUTp-1l, SpMin6_Bhe, SpMin8_Bhp and IC2) were negatively correlated with the PNECs.

All the influencing molecular descriptors are two-dimensional molecular descriptors, including autocorrelation descriptors, information index descriptors, burden modified eigenvalues descriptors, barysz matrix descriptors, and BCUT descriptors. The effects of these molecular descriptors on the biotoxicity had been reported before and the developed QSAR model was also used for the estimation of the toxicity concentration of some chemicals [54,60,61]. For instance, the autocorrelation descriptors (GATS7p, MATS1p, ATSC5v, MATS8e, ATSC2p, ATSC1m), the burden modified eigenvalues descriptors (SpMax2_Bhp, SpMin4_Bhe, SpMin2_Bhs, SpMin1_Bhs), and the BCUT descriptor (BCUTw-1h), were observed to significantly affect the interspecies toxicity of 1,2,4-triazole compounds to mice [24]. The BCUT descriptors and the information index descriptors were investigated as important molecular descriptors on the toxicity of alcohol compounds to *Rana temporaria* [62]. Three autocorrelation parameters (GATS5s, GATS1p, and ATSC7v) and the barysz matrix descriptor (VE3 DzZ) were useful in estimating the acute toxicity of the emerging contaminants such as active ingredients and their metabolites, ingredients of cosmetic and personal care products, pesticides and their trans-formation products to freshwater invertebrates [25]. The autocorrelation parameters (ATSC2e, MATS2v, ATSc2, MATS6s), the BCUT descriptor (BCUTw-1l), and the burden modified eigenvalues descriptor (SpMax5_Bhs) were used for estimating the acute oral toxicity of PAHs to mammals by a two-dimensional parametric model using genetic algorithms and multiple linear regression [31]. However, the quantitative relationships between the molecular structure and PNECs and the application of these quantitative relationships to estimate the PNECs are rarely reported.

As shown in Fig. 4, the influencing weight of these molecular descriptors contributed to the PNECs in each model was used to characterize the influence of the molecular descriptors on the PNECs. For the developed three models, the autocorrelation descriptors (e.g., ATSC4i, MATS2c, GATS3c), information index descriptors (e.g., IC2, ZMIC4, CIC5), burden modified eigenvalues descriptors (e.

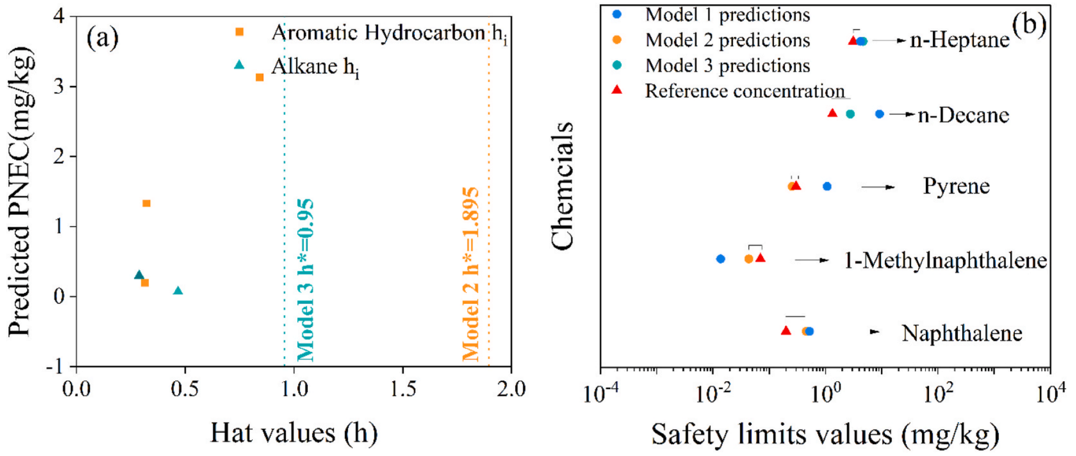


Fig. 5. (a) The defined application domains range of the models; (b)Comparison of the estimated PNECs and the proposed safety limits of the PHDs.

g., SpMax3_Bhs, SpMin1_Bhp), barysz matrix descriptors (e.g., VR3_D, VE3_DzZ), and BCUT descriptors (BCUTp-1l), were accounted for 45.67%, 22.13%, 16.67%, 8.17%, and 7.47% of the weight in all the influencing molecular descriptors.

3.6. The mechanism underlying the quantitative relationships

The quantitative relationships between the molecular structure and PNECs indicated that 34 two-dimensional molecular descriptors, including autocorrelation descriptors, information index descriptors, burden modified eigenvalues descriptors, barysz matrix descriptors, and BCUT descriptors, were associated with the toxicity of PHDs. The obtained results can provide insights into the underlying mechanisms for the effects of these molecular descriptors on the toxicity and PNECs of PHDs.

Autocorrelation descriptors are important molecular descriptors affecting the PNECs and biotoxicity of AHDs in this study. Among the 34 influencing molecular descriptors, 18 molecular descriptors (ATSC2c, ATSC2s, ATSC2v, ATSC4e, ATSC4i, ATSC4c, AATSC4v, AATSC2c, AATSC1e, AATSC2m, AATSC4p, MATS4v, MATS1p, MATS2c, GATS3c, GATS3i, GATS4c, and GATS1s) were autocorrelation descriptors in this study. A high proportion of the autocorrelation descriptors showed significant effects on the PNECs of PHDs. The autocorrelation descriptors that affected the PNECs of PHDs were mainly the mass (m), polarizability (p), van der Waals volume (v), first ionization potential (i), and state (s) weighting of Broto-Moreau (AST), Geary (GAT), and Moran (MATS) descriptors. The three autocorrelation descriptors characterized the structural conformation of chemical molecules [49,63], which has been found to be highly relevant to the aquatic toxicity of cosmetics and personal care additives [49]. These autocorrelation descriptors may affect the PNECs and toxicity by influencing the spatial conformation of PHDs. The specific weighting of the autocorrelation descriptors such as the mass (m), polarizability (p), van der Waals volume (v), first ionization potential (i), and state (s), which characterized the functions and properties of atoms in a molecule, were also found to be important factors on the PNECs and toxicity of PHDs. Taking the polarizability, mass, and first ionization potential weighting of the autocorrelation descriptors as examples. AATSC4p and MATS1p are autocorrelation descriptors weighted by atomic polarizability, describing the overall mobility of electrons and the reactivity of a chemical, accounted for 10.4% and 5.9% of the weight in all the influencing molecular descriptors (Fig. 4). The great influence of AATSC4p and MATS1p on the PNECs and toxicity of PHDs is probably affected by the atomic polarizability. A chemical with a high polarization is usually not easy to cross the biofilm to accumulate in biological tissues, generally resulting in its low toxicity [25]. This is consistent with the results in this study that AATSC4p and MATS1p are positively relevant to the PNEC and negatively related with the toxicity of PHDs. AATSC2m is an autocorrelation descriptor weighted by atomic mass, measuring the strength between relative atomic mass of the atom pairs, accounted for 9.6% in Model 2 and 4.2% in Model 3 of the weight in all the influencing molecular descriptors. The positive correlation between AATSC2m and PNEC might be influenced by the atomic mass. A chemical with a greater molecular mass is usually more difficult to enter into the organisms and then act on the active site [64] and thus produces less toxic effects [65]. GATS3i is a 2D Geary autocorrelation descriptor weight by the first ionization potential of atom pairs, describing the ionization potential from the molecules with several carbon—carbon bonds, accounted for 3.8% of the weight in all the influencing molecular descriptors. Molecular with a lower GATS3i value usually has a higher carbon content which might lead to a higher toxic effect [66].

Individual information index descriptors were observed to the most influencing molecular descriptors on the PNECs of PHDs in the present study. The complementary information content index of the neighborhood symmetry of order-5 (CIC5) and the information content index of the neighborhood symmetry of order-2 (IC2) showed the maximum influencing weight (17.7% and 31.5%) on the PNECs of AHDs and ALKDs, respectively. IC2 also showed a high negative contribution to the PNECs of all the PHDs, with a high influencing weight of 9.2%. Many studies have focused on the relationship between the information index descriptors and the toxicity, however, the correlation between the information index descriptors and the PNECs is still not clear. Taking IC2 as examples, IC2 primarily represents the topological features and information transfer capabilities of chemical molecules [67]. A higher IC2 indicates a stronger information transfer among the atoms within the molecule and a higher molecular connectivity of a chemical. As a result, the chemical appeared to exhibit a higher diffusion coefficient and a stronger interaction, and thus potentially presented a greater toxic effect [68]. Therefore, the positive correlation between IC2 and biotoxicity was obviously observed in the PHDs with longer molecular topological distances.

Burden modified eigenvalues descriptors and barysz matrix descriptors, derived from the Burden and Barysz matrices, were also important in influencing the PNECs and toxicity of PHDs in this study. Six Burden modified eigenvalues descriptors (SpMax3_Bhs, SpMin1_Bhp, SpMin6_Bhs, SpMin6_Bhi, SpMin8_Bhp, SpMin6_Bhe) and barysz matrix descriptors (VE2_Dzs, VR3_D, VE3_DzZ, VE3_Dzs), showed a 16.67% and 8.17% weight in all the influencing molecular descriptors, respectively. The two types of molecular descriptors were related to the molecular topological characteristics of chemicals that associated with the molecular size, the atomic number, and the content of some specific heteroatoms with a role in the toxicity [25,29], and thus affected the PNECs and toxicity of PHDs.

The first lowest eigenvalue in the Burden matrix weighted by polarizability (BCUTp-1l) was a significant BCUT descriptor affecting the PNECs and toxicity of ALKDs in this study. BCUTp-1l was negatively related with the PNECs of ALKDs, contributing 22.4% weight to the PNECs in the Model 3. Previous studies have reported that high BCUTp-1l demonstrated a higher spatial metric polarizability that describing electron mobility and reactivity of a chemical, and thus resulted in higher activity and toxic effects on organisms [69], which was in agreement with the results in this study.

The list of the abbreviations and its definition in this study.

Abbreviations	Definition
ESTs	Environmental safety thresholds
QSAR	Quantitative structure-activity relationship
PNECs	Predicted no-effect concentrations
PHDs	Petroleum hydrocarbons and their derivatives
AHDs	Aromatic hydrocarbons and their derivatives
ALKDs	Alkanes, Alkenes, and their derivatives
TPHs	Total petroleum hydrocarbons
SSD	Species sensitivity distribution
AF	Assessment factor
AHs	Aromatic hydrocarbons
EC_{50}	Median effective concentration
LD_{50}	Median lethal dose
NOEC	Chronic toxicity concentrations
LOEC	Lowest observed effect concentration
NOEL	No observed effect level
$L(E)C_{10}$	The concentration causing a 10% effect within a specified time interval
MLR	Multiple linear regression
VIF	Variance inflation factor
AD	Application domain

4. Conclusions

(1) Three validated QSAR models with high accuracy in the estimation of PNECs were separately developed for PHDs, AHDs and ALKDs. The separate model developed for AHDs and ALKDs showed better performance in estimating the PNECs.

(2) The developed QSAR models showed wide application domain range, supporting a new cost-effective and reliable approach for directly estimating the PNECs of PHDs in the ecological risk assessment.

(3) 34 two-dimensional molecular descriptors were observed to influence the PNECs of PHDs. Most of the involved molecular descriptors were autocorrelation descriptors, and the individual information index descriptors contributed the highest weight in all the influencing molecular descriptors on the PNECs of PHDs.

(4) The quantitative relationships between the molecular descriptors and PNECs provides new insights into understanding the mechanism of the effects of the associated molecular descriptors on the toxicity and PNECs of PHDs.

Data availability statement

The authors confirm that the data supporting the findings of this study are available within the manuscript and its supplementary materials.

CRediT authorship contribution statement

Jiajia Wei: Writing – original draft, Validation, Investigation, Formal analysis, Data curation, Conceptualization. **Lei Tian:** Writing – review & editing, Validation, Methodology, Investigation, Funding acquisition, Formal analysis. **Fan Nie:** Validation, Project administration, Methodology, Formal analysis. **Zhiguo Shao:** Validation, Supervision, Project administration, Methodology. **Zhansheng Wang:** Validation, Supervision, Project administration. **Yu Xu:** Validation, Project administration, Methodology, Formal analysis. **Mei He:** Writing – review & editing, Validation, Supervision, Project administration, Investigation, Funding acquisition, Formal analysis.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This study was supported by the CNPC Scientific Research and Technology Development Programme (2021DJ6605), and the Yangtze Talents Fund [2020–2023].

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.heliyon.2024.e26808.

References

[1] B. Wu, S. Guo, J. Wang, Spatial ecological risk assessment for contaminated soil in oiled fields, J. Hazard Mater. 403 (2021) 123984, https://doi.org/10.1016/j.jhazmat.2020.123984.

[2] J. Nishiwaki, Y. Kawabe, Y. Sakamoto, et al., Volatilization properties of gasoline components in soils, Environ. Earth Sci. 63 (2011) 87–95, https://doi.org/10.1007/s12665-010-0671-7.

[3] B. Wu, S. Guo, L. Zhang, et al., Spatial variation of residual total petroleum hydrocarbons and ecological risk in oilfield soils, Chemosphere 291 (2021) 132916, https://doi.org/10.1016/j.chemosphere.2021.132916.

[4] Q. Liu, C. Xia, L. Wang, et al., Fingerprint analysis reveals sources of petroleum hydrocarbons in soils of different geographical oilfields of China and its ecological assessment, Sci. Rep. 12 (1) (2022) 4808, https://doi.org/10.1038/s41598-022-08906-6.

[5] M. Andrade-Couce, P. Marcet, L. Fernández-Feal, et al., Impact of the Prestige oil spill marsh soils: relationship between heavy metal, sulfide and total petroleum hydrocarbon contents at the Villarrube and Lires marshes (Galicia, Spain), Cienc. Mar. 30 (2004) 477–487, https://doi.org/10.7773/cm.v30i3.281.

[6] Standardization Administration of the people's Republic of China, Soil Environmental Quality-Risk Control Standard for Soil Contamination of Development Land, GB36600-2018, 2015. Beijing, China, https://www.mee.gov.cn/ywgz/fgbz/bz/bzwb/trhj/201807/t20180703_446027.shtml.

[7] M. Alkio, T.M. Tabuchi, X. Wang, et al., Stress responses to polycyclic aromatic hydrocarbons in Arabidopsis include growth inhibition and hypersensitive response-like symptoms, J. Exp. Bot. 56 (421) (2005) 2983–2994, https://doi.org/10.1093/jxb/eri295.

[8] L. Wang, B. Zheng, W. Meng, Photo-induced toxicity of four polycyclic aromatic hydrocarbons, singly and in combination, to the marine diatom Phaeodactylum tricornutum, Ecotoxicol. Environ. Saf. 71 (2) (2008) 465–472, https://doi.org/10.1016/j.ecoenv.2007.12.019.

[9] J. Wen, L. Pan, Short-term exposure to benzo[a]pyrene causes oxidative damage and affects haemolymph steroid levels in female crab Portunus trituberculatus, Environ. Pollut. 208 (2016) 486–494, https://doi.org/10.1016/j.envpol.2015.10.019.

[10] L. Yang, W.-C. Wang, S.-C.C. Lung, et al., Polycyclic aromatic hydrocarbons are associated with increased risk of chronic obstructive pulmonary disease during haze events in China, Sci. total environ. 574 (2017) 1649–1658, https://doi.org/10.1016/j.scitotenv.2016.08.211.

[11] J.A. Song, C.Y. Choi, Exposure to benzo[α]pyrene causes oxidative stress and cell damage in bay scallop Argopecten irradians, Aquac rep 21 (2021) 100860, https://doi.org/10.1016/j.aqrep.2021.100860.

[12] H. Wang, X. Huang, Z. Kuang, et al., Source apportionment and human health risk of PAHs accumulated in edible marine organisms: a perspective of "source-organism-human", J. Hazard Mater. 453 (2023) 131372 https://doi.org/10.1016/j.jhazmat.2023.131372.

[13] H. Han, S. Huang, S. Liu, et al., An assessment of marine ecosystem damage from the penglai 19-3 oil spill accident, J. Mar. Sci. Eng. 9 (2021) 732, https://doi.org/10.3390/jmse9070732.

[14] B. Andres, The exxon valdez oil spill disrupted the breeding of black oystercatchers, J Wildl 61 (1997) 1322, https://doi.org/10.2307/3802132.

[15] M.I. Khan, S.A. Cheema, X. Tang, et al., A battery of bioassays for the evaluation of phenanthrene biotoxicity in soil, Arch. Environ. Contam. Toxicol. 65 (1) (2013) 47–55, https://doi.org/10.1007/s00244-013-9879-3.

[16] C. Russom, R. Breton, J. Walker, et al., An overview of the use of quantitative structure-activity relationships for ranking and prioritizing large chemical inventories for environmental risk assessments, Environ. Toxicol. Chem. 22 (2003) 1810–1821, https://doi.org/10.1897/01-194.

[17] S. Tao, X. Xi, F. Xu, et al., A fragment constant QSAR model for evaluating the EC_{50} values of organic chemicals to Daphnia magna, Environ. Pollut. 116 (1) (2002) 57–64, https://doi.org/10.1016/S0269-7491(01)00119-1.

[18] S. Samanipour, J.W. O'Brien, M.J. Reid, et al., From molecular descriptors to intrinsic fish toxicity of chemicals: an alternative approach to chemical prioritization, Environ. Sci. Technol. (2022), https://doi.org/10.1021/acs.est.2c07353.

[19] A.A. Toropov, M.R. Di Nicola, A.P. Toropova, et al., A regression-based QSAR-model to predict acute toxicity of aromatic chemicals in tadpoles of the Japanese brown frog (Rana japonica): calibration, validation, and future developments to support risk assessment of chemicals in amphibians, Sci. Total Environ. 830 (2022) 154795, https://doi.org/10.1016/j.scitotenv.2022.154795.

[20] A. Golbamaki, A. Cassano, A. Lombardo, et al., Comparison of in silico models for prediction of Daphnia magna acute toxicity, SAR QSAR Environ. Res. 25 (8) (2014) 673–694, https://doi.org/10.1080/1062936x.2014.923041.

[21] M. Cassotti, V. Consonni, A. Mauri, et al., Validation and extension of a similarity-based approach for prediction of acute aquatic toxicity towards Daphnia magna, SAR QSAR Environ. Res. 25 (12) (2014) 1013–1036, https://doi.org/10.1080/1062936x.2014.977818.

[22] K.P. Singh, S. Gupta, A. Kumar, et al., Multispecies QSAR modeling for predicting the aquatic toxicity of diverse organic chemicals for regulatory toxicology, Chem. Res. Toxicol. 27 (5) (2014) 741–753, https://doi.org/10.1021/tx400371w.

[23] X. Yu, Q. Zeng, Random forest algorithm-based classification model of pesticide aquatic toxicity to fishes, Aquat. Toxicol. 251 (2022) 106265, https://doi.org/10.1016/j.aquatox.2022.106265.

[24] Z. Liu, K. Dang, J. Gao, et al., Toxicity prediction of 1,2,4-triazoles compounds by QSTR and interspecies QSTTR models, Ecotoxicol. Environ. Saf. 242 (2022) 113839, https://doi.org/10.1016/j.ecoenv.2022.113839.

[25] G.J. Lavado, D. Baderna, D. Gadaleta, et al., Ecotoxicological QSAR modeling of the acute toxicity of organic compounds to the freshwater crustacean Thamnocephalus platyurus, Chemosphere 280 (2021) 130652, https://doi.org/10.1016/j.chemosphere.2021.130652.

[26] E. Zvinavashe, T. Du, T. Griff, et al., Quantitative structure-activity relationship modeling of the toxicity of organothiophosphate pesticides to Daphnia magna and cyprinus carpio, Chemosphere 75 (11) (2009) 1531–1538, https://doi.org/10.1016/j.chemosphere.2009.01.081.

[27] X. Li, T. Zhang, X. Min, et al., Toxicity of aromatic compounds to Tetrahymena estimated by microcalorimetry and QSAR, Aquat. Toxicol. 98 (4) (2010) 322–327, https://doi.org/10.1016/j.aquatox.2010.03.002.

[28] W. Gu, X. Li, M. Du, et al., Identification and regulation of ecotoxicity of polychlorinated naphthalenes to aquatic food Chain (green algae-Daphnia magna-fish), Aquat. Toxicol. 233 (2021) 105774, https://doi.org/10.1016/j.aquatox.2021.105774.

[29] S. Chen, G. Sun, T. Fan, et al., Ecotoxicological QSAR study of fused/non-fused polycyclic aromatic hydrocarbons (FNFPAHs): assessment and priority ranking of the acute toxicity to Pimephales promelas by QSAR and consensus modeling methods, Sci. Total Environ. 876 (2023) 162736, https://doi.org/10.1016/j.scitotenv.2023.162736.

[30] Y. Zhang, Y. Zhu, Y. Shao, et al., Toxicity of disinfection byproducts formed during the chlorination of sulfamethoxazole, norfloxacin, and 17β-estradiol in the presence of bromide, Environ. Sci. Pollut. Res. 28 (36) (2021) 50718–50730, https://doi.org/10.1007/s11356-021-14161-5.

[31] G. Sun, Y. Zhang, L. Pei, et al., Chemometric QSAR modeling of acute oral toxicity of Polycyclic Aromatic Hydrocarbons (PAHs) to rat using simple 2D descriptors and interspecies toxicity modeling with mouse, Ecotoxicol. Environ. Saf. 222 (2021) 112525, https://doi.org/10.1016/j.ecoenv.2021.112525.

[32] Y. Wang, J. Wang, J. Mu, et al., Aquatic predicted no-effect concentration for three polycyclic aromatic hydrocarbons and probabilistic ecological risk assessment in Liaodong Bay of the Bohai Sea, China, Environ. Sci. Pollut. Res. Int. 21 (1) (2014) 148–158, https://doi.org/10.1007/s11356-013-1597-x.

[33] L. Zeng, S. Zeng, X. Dong, et al., Probabilistic ecological risk assessment of polycyclic aromatic hydrocarbons in southwestern catchments of the Bohai Sea, China, Ecotoxicology 22 (8) (2013) 1221–1231, https://doi.org/10.1007/s10646-013-1110-9.

[34] K.M.Y. Leung, J.S. Gray, W.K. Li, et al., Deriving sediment quality guidelines from field-based species sensitivity distributions, Environ. Sci. Technol. 39 (14) (2005) 5148–5156, https://doi.org/10.1021/es050450x.

[35] J.-K. Im, Y.-C. Cho, H.-R. Noh, et al., Geographical distribution and risk assessment of volatile organic compounds in tributaries of the han river watershed, Agronomy 11 (2021) 956, https://doi.org/10.3390/agronomy11050956.

[36] X. Jin, J. Zha, Y. Xu, et al., Derivation of predicted no effect concentrations (PNEC) for 2,4,6-trichlorophenol based on Chinese resident species, Chemosphere 86 (1) (2012) 17–23, https://doi.org/10.1016/j.chemosphere.2011.08.040.

[37] A. Finizio, S. Villa, M. Vighi, Predicted No effect concentration (PNEC), Reference module in biomedical Sciences (2021), https://doi.org/10.1016/B978-0-12-824315-2.00004-X.

[38] H. Abdel-Shafy, M. Mansour, A review on polycyclic aromatic hydrocarbons: source, environmental impact, effect on human health and remediation, Egypt J Pet 25 (2015) 107–123, https://doi.org/10.1016/j.ejpe.2015.03.011.

[39] R.M. Flores-Serrano, R. Iturbe-Argüelles, G. Pérez-Casimiro, et al., Ecological risk assessment for small omnivorous mammals exposed to polycyclic aromatic hydrocarbons: a case study in northeastern Mexico, Sci. Total Environ. 476–477 (2014) 218–227, https://doi.org/10.1016/j.scitotenv.2013.12.092.

[40] H. Fan, Y. Wang, X. Liu, et al., Derivation of predicted no-effect concentrations for thirty-five pharmaceuticals and personal care products to freshwater ecosystem, Front. Mar. Sci. 9 (2022) 1043792, https://doi.org/10.3389/fmars.2022.1043792.

[41] D.T. Salvito, R.J. Senna, T.W. Federle, A framework for prioritizing fragrance materials for aquatic risk assessment, Environ. Toxicol. Chem. 21 (6) (2002) 1301–1308, https://doi.org/10.1002/etc.5620210627.

[42] Y. Chen, X. Xi, G. Yu, et al., Pharmaceutical compounds in aquatic environment in China: locally screening and environmental risk assessment, Front. Environ. Sci. 9 (3) (2015) 394–401, https://doi.org/10.1007/s11783-014-0653-1.

[43] K. Sorgog, M. Kamo, Quantifying the precision of ecological risk: conventional assessment factor method vs. species sensitivity distribution method, Ecotoxicol. Environ. Saf. 183 (2019) 109494, https://doi.org/10.1016/j.ecoenv.2019.109494.

[44] A.I. Okonski, D.B. MacDonald, K. Potter, et al., Deriving predicted no-effect concentrations (PNECs) using a novel assessment factor method, Hum. Ecol. Risk Assess. 27 (6) (2021) 1613–1635, https://doi.org/10.1080/10807039.2020.1865788.

[45] Agency, U.S.E.P, Short-term methods for estimating the chronic toxicity of effluents and receiving waters to freshwater organisms, in: 821-R-02-013, 2002.

[46] M. Hamadache, O. Benkortbi, S. Hanini, et al., A Quantitative structure activity relationship for acute oral toxicity of pesticides on rats: validation, domain of application and prediction, J. Hazard Mater. 303 (2016) 28–40, https://doi.org/10.1016/j.jhazmat.2015.09.021.

[47] Z. Cai, M. Zafferani, O.M. Akande, et al., Quantitative structure–activity relationship (QSAR) study predicts small-molecule binding to RNA structure, J. Med. Chem. 65 (10) (2022) 7262–7277, https://doi.org/10.1021/acs.jmedchem.2c00254.

[48] A. Golbraikh, M. Shen, Z. Xiao, et al., Rational selection of training and test sets for the development of validated QSAR models, J. Comput. Aided Mol. Des. 17 (2–4) (2003) 241–253, https://doi.org/10.1023/a:1025386326946.

[49] Y.-T. Yang, H.-G. Ni, Predictive in silico models for aquatic toxicity of cosmetic and personal care additive mixtures, Water Res. 236 (2023) 119981, https://doi.org/10.1016/j.watres.2023.119981.

[50] Z. Fang, X. Yu, Q. Zeng, Random forest algorithm-based accurate prediction of chemical toxicity to Tetrahymena pyriformis, Toxicology 480 (2022) 153325, https://doi.org/10.1016/j.tox.2022.153325.

[51] B.K. Sharma, P. Pilania, P. Singh, et al., CP-MLR directed QSAR study of carbonic anhydrase inhibitors: sulfonamide and sulfamate inhibitors, Cent. Eur. J. Chem. 7 (4) (2009) 909–922, https://doi.org/10.2478/s11532-009-0073-4.

[52] K. Roy, S. Kar, P. Ambure, On a simple approach for determining applicability domain of QSAR models, Chemometr. Intell. Lab. Syst. 145 (2015) 22–29, https://doi.org/10.1016/j.chemolab.2015.04.013.

[53] X. Wu, J. Guo, G. Dang, et al., Prediction of acute toxicity to Daphnia magna and interspecific correlation: a global QSAR model and a Daphnia-minnow QTTR model, SAR QSAR Environ. Res. 33 (8) (2022) 583–600, https://doi.org/10.1080/1062936x.2022.2098814.

[54] T. Bo, Y. Lin, J. Han, et al., Machine learning-assisted data filtering and QSAR models for prediction of chemical acute toxicity on rat and mouse, J. Hazard Mater. 452 (2023) 131344, https://doi.org/10.1016/j.jhazmat.2023.131344.

[55] F. Sahigara, K. Mansouri, D. Ballabio, et al., Comparison of different approaches to define the applicability domain of QSAR models, Molecules (2012) 4791–4810, https://doi.org/10.3390/molecules17054791.

[56] M. Grzonkowska, A. Sosnowska, M. Barycki, et al., How the structure of ionic liquid affects its toxicity to Vibrio fischeri? Chemosphere 159 (2016) 199–207, https://doi.org/10.1016/j.chemosphere.2016.06.004.

[57] H. Liu, M. Wei, X. Yang, et al., Development of TLSER model and QSAR model for predicting partition coefficients of hydrophobic organic chemicals between low density polyethylene film and water, Sci. Total Environ. 574 (2016), https://doi.org/10.1016/j.scitotenv.2016.08.051.

[58] J.M. Bornstein, J. Adams, B. Hollebone, et al., Effects-driven chemical fractionation of heavy fuel oil to isolate compounds toxic to trout embryos, Environ. Toxicol. Chem. 33 (4) (2014) 814–824, https://doi.org/10.1002/etc.2492.

[59] J.A. McGrath, D.M. Di Toro, Validation of the target lipid model for toxicity assessment of residual petroleum constituents: monocyclic and polycyclic aromatic hydrocarbons, Environ. Toxicol. Chem. 28 (6) (2009) 1130–1148, https://doi.org/10.1897/08-271.1.

[60] S. Chen, G. Sun, T. Fan, et al., Ecotoxicological QSAR study of fused/non-fused polycyclic aromatic hydrocarbons (FNFPAHs): assessment and priority ranking of the acute toxicity to Pimephales promelas by QSAR and consensus modeling methods, Sci. Total Environ. 876 (2023) 162736, https://doi.org/10.1016/j.scitotenv.2023.162736.

[61] W. Di Marzio, M.E. Saenz, Quantitative structure-activity relationship for aromatic hydrocarbons on freshwater fish, Ecotoxicol. Environ. Saf. 59 (2) (2004) 256–262, https://doi.org/10.1016/j.ecoenv.2003.11.006.

[62] L. Wang, P. Xing, C. Wang, et al., Maximal information coefficient and support vector regression based nonlinear feature selection and QSAR modeling on toxicity of alcohol compounds to tadpoles of Rana temporaria, J. Braz. Chem. Soc. 30 (2019) 279–285, https://doi.org/10.21577/0103-5053.20180176.

[63] L. Yang, R. Tian, Z. Li, et al., Data driven toxicity assessment of organic chemicals against Gammarus species using QSAR approach, Chemosphere 328 (2023) 138433, https://doi.org/10.1016/j.chemosphere.2023.138433.

[64] T. Rezić, A. Vrsalović Presečki, Ž. Kurtanjek, New approach to the evaluation of lignocellulose derived by-products impact on lytic-polysaccharide monooxygenase activity by using molecular descriptor structural causality model, Bioresour. Technol. 342 (2021) 125990, https://doi.org/10.1016/j.biortech.2021.125990.

[65] S. Adawara, G. Shallangwa, P. Mamza, et al., Molecular docking and QSAR theoretical model for prediction of phthalazinone derivatives as new class of potent dengue virus inhibitors, Beni-Suef univ j basic appl sci. 9 (2020), https://doi.org/10.1186/s43088-020-00073-9.

[66] M. Cassotti, D. Ballabio, R. Todeschini, et al., A similarity-based QSAR model for predicting acute toxicity towards the fathead minnow (Pimephales promelas), SAR QSAR Environ. Res. 26 (6) (2015) 521, https://doi.org/10.1080/1062936x.2015.1035056.

[67] R. Yijun, X. Xiao-Ke, J. Tao, The Maximum Capability of a Topological Feature in Link Prediction, 2022, https://doi.org/10.48550/arXiv.2206.15101.

[68] T. Zhu, Y. Jiang, H. Cheng, et al., Development of pp-LFER and QSPR models for predicting the diffusion coefficients of hydrophobic organic compounds in LDPE, Ecotoxicol. Environ. Saf. 190 (2020) 110179, https://doi.org/10.1016/j.ecoenv.2020.110179.

[69] R.S. Pearlman, K.M. Smith, Novel software tools for chemical diversity, Perspect. Drug Discov. Des. 9 (1998) 339–353, https://doi.org/10.1023/A:1027232610247.